DigitaLung

Digitales Auskultationssystem zur Differentialdiagnose von Lungenerkrankungen mittels Machine Learning

Teilvorhaben: Datenerarbeitung und Maschinelles Lernen
(Förderkennzeichen: 13GW0554D)

Abschlussbericht

Verbundvorhaben:	**Digitales Auskultationssystem zur Differentialdiagnose von Lungenerkrankungen mittels Machine Learning**
Projektleitung (LUH):	**Prof. Dr. Ing. Bodo Rosenhahn** **Institut für Informationsverarbeitung** **Leibniz Universität Hannover** **30167 Hannover**
Laufzeit des Projektes:	**01.11.2021 - 30.04.2025**

Am Vorhaben beteiligte Institute/Einrichtungen:

- ERKA Kallmeyer Medizintechnik GmbH & Co. KG
  Björn Kallmeyer (Koordination, Leitung)
- Fraunhofer-Institut für Toxikologie und Experimentelle Medizin (ITEM)
  Dr. Sylvia Escher
- Medizinische Hochschule Hannover (MHH)
  PD Dr. Jessica Radema-cher
- Leibniz Universität Hannover (LUH)
  Prof Dr. Bodo Rosenhahn

Projektförderer/Projektträger

Das Verbundvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 13GW0554D gefördert. Die Verantwortung für die Inhalte des Schlussberichts liegt bei den jeweils genannten Autoren.

Inhaltsverzeichnis

Abbildungsverzeichnis **4**

1 Kurzdarstellung **5**
1.1 Aufgabenstellung 5
1.2 Stand von Wissenschaft und Technik 5
1.3 Ablauf des Vorhabens und wesentliche Ergebnisse 6
1.4 Zusammenarbeit mit anderen Stellen 6

2 Eingehende Darstellung **7**
2.1 Verwendung der Zuwendung 7
2.1.1 Patientenrekrutierung und Datensammlung (AP3) 7
2.1.2 Vorbereitende Datenanalyse (AP4) 8
2.1.3 Datenverarbeitung und Maschinelles Lernen (AP5) 11
2.1.4 Datenaufbereitung 11
2.1.5 Erweiterte ML-Untersuchungen (AP9) 14
2.1.6 Sonstige Arbeitspakete 16
2.2 Notwendigkeit und Angemessenheit der geleisteten Arbeit 17
2.3 Voraussichtlicher Nutzen 18
2.4 Erfolgte Veröffentlichungen 19

Literaturverzeichnis **20**

Abbildungsverzeichnis

2.1 Interface von Labelstudio 8
2.2 Visualisierung einiger ungewollter Signaleigenschaften im Jordan-Datensatz. Dazu gehören: zu leises Atmen (b), Mikrofonreibung (c) am Körper und zu lautes Atmen (d). 15
2.3 Visualisierung der Grad-CAM Ergebnisse zwischen Klasse Gesund (a) und Asthma (b). 16
2.4 Visualisierung der Modellaktivierung für Asthma-Artefakte (Grad-CAM). Die Rote Box markiert den Bereich, den das Modell vorzugsweise als Feature für die klasse Asthma identifiziert 17

1 Kurzdarstellung

1.1 Aufgabenstellung

Als zentraler Bestandteil der klinischen Untersuchung ermöglicht die Auskultation der Lungengeräusche mittels Stethoskop die Erkennung pulmonaler Erkrankungen. Ihr Hauptproblem liegt jedoch in ihrer Subjektivität: Die Beurteilung hängt stark vom individuellen Hörvermögen und der persönlichen Erfahrung des untersuchenden Facharztes ab. Das Ziel des Verbundprojekts *DigitaLung* besteht darin, die Machbarkeit eines fortschrittlichen Machine Learning-gestützten digitalen Auskultationssystems zu demonstrieren, um weniger erfahreneren Ärztinnen und Ärzten bei der Diagnose zu unterstützen und die Gesundheitsversorgung von Patienten nachhaltig zu verbessern. Das System nutzt multimodale Sensortechnik, um die Diagnose von Atemwegserkrankungen wie Bronchitis, Pneumonie, Asthma, COPD und COVID-19 zu verbessern.

1.2 Stand von Wissenschaft und Technik

Die Grundlage für die computergestützte Analyse von Atemgeräuschen entstand bereits in den 1970er Jahren [6]. Dieses Verfahren ermöglicht eine qualitative und quantitative Erfassung der akustischen Lungensignale, welche sich über ein Frequenzband von 50 bis 2000 Hz erstrecken. Man unterscheidet hierbei zwischen den normalen Hauptgeräuschen sowie kontinuierlichen (z.B. Pfeifen) und diskontinuierlichen (z.B. Rasseln) Nebengeräuschen, deren Zuordnung zu Krankheitsbildern gut dokumentiert ist [2]. Wichtige Standards für die technische Umsetzung etablierte die europäische CORSA-Studie im Jahr 2000 [4]. Die Gerätebasis umfasst Sensoren, eine Verstärker- und Filtereinheit sowie einen A/D-Wandler. Das zentrale digitale Untersuchungsverfahren ist die Spektralanalyse, bei der die akustischen Signale mithilfe der Fourier- oder Wavelet-Transformation in die Frequenzdomäne (Spektrogramm) überführt werden, um spektrale Eigenschaften zu quantifizieren [2].

1.3 Ablauf des Vorhabens und wesentliche Ergebnisse

Das Forschungsvorhaben *DigitaLung* umfasst die Entwicklung eines multimodalen Stethoskops mit zehn Audio-Kanälen zur präziseren Klassifikation von Lungenkrankheiten. Dieses hochspezialisierte Stethoskop bildet die Voraussetzung für die Durchführung einer Patientenstudie, mit der ein eigener Datensatz an Lungengeräuschen erstellt werden soll. Die so gewonnenen Daten dienen als Basis für das Training der KI-gestützten Auswerteeinheit. Aufgrund zunehmender Verzögerungen in der Datensatzakquise wurde die Entscheidung getroffen, das KI-Modell auf dem öffentlich verfügbaren Jorden-Datensatz zu entwickeln. Das daraus resultierende Modell erreicht eine Klassifikationsgenauigkeit (Accuracy) von 67% bei einem Recall-Wert von 69%.

Um die Entscheidungsfindung des Modells transparenter zu gestalten, wurde zudem das interpretierbare Grad-CAM-Methode implementiert. Dieses Verfahren ermöglicht die Identifikation jener Geräuschmuster, welche maßgeblich für oder gegen eine spezifische diagnostische Klassifikation sprechen.

1.4 Zusammenarbeit mit anderen Stellen

Die Koordination mit den beteiligten Projektgruppen stützte sich auf regelmäßige, Status-Meetings, welche als zentrale Plattform dienten, um den jeweiligen Projektfortschritt transparent zu teilen und kritische Schnittstellen oder Abhängigkeiten frühzeitig zu identifizieren. Als verantwortliche Teilgruppe für die Entwicklung der zentralen Auswerteeinheit (LUH) übernahmen wir eine Schlüsselrolle bei der Definition präziser Datensatzanforderungen. Zudem leisteten wir aktive technische Unterstützung bei der Integration und Anpassung des Labeling-Tools für die Datenannotation.

Zusätzlich zum Treffen der Gesamtgruppe fand eine engere Abstimmung in der KI-Fachgruppe statt, die sich aus Experten der MHH und der LUH zusammensetzte. In diesen fokussierten Treffen wurden spezifische Herausforderungen der Datenaufbereitung (wie Datenvorverarbeitung und Optimierungsverfahren) sowie die Evaluierung und Abstimmung der methodischen Modellansätze (Algorithmus, Design und Architektur) intensiv diskutiert, um die technische und inhaltliche Kohärenz der KI-Komponenten sicherzustellen.

2 Eingehende Darstellung

Dieses Kapitel bietet einen detaillierten Überblick über die im Projekt erzielten Ergebnisse und dessen Kontext. Die Zuwendungsnutzung wird zunächst anhand der Arbeitspakete (AP3 bis AP9) in Kapitel 2.1 präsentiert, mit Hauptaugenmerk auf der Datensammlung und -verarbeitung, einschließlich vorbereitender Analyse zur ML-Anwendungen. Es folgt die Beschreibung der Integration des ML-Diagnosesystems und der erzielten Ergebnisse. Anschließend werden die Erweiterung des Systems um Interpretable AI (erklärbare KI) sowie die damit verbundenen Ergebnisse präsentiert. In Kapitel 2.2 und 2.3 werden die Notwendigkeit und Angemessenheit der geleisteten Arbeiten bewertet, sowie der voraussichtliche Nutzen des Endprodukts herausgestellt.

2.1 Verwendung der Zuwendung

In den folgenden Abschnitten erfolgt eine detaillierte Darstellung des Projektfortschritts und der spezifischen wissenschaftlich-technischen Ergebnisse Die Beschreibung ist systematisch nach den definierten Arbeitspaketen (APs) gegliedert. Es ist anzumerken, dass sich die Inhalte der Arbeitspakete teilweise überschneiden. Die Gliederung dient daher lediglich als grober Ordnungsrahmen zur inhaltlichen Orientierung.

2.1.1 Patientenrekrutierung und Datensammlung (AP3)

Im Rahmen der Datensammlung hat die LUH maßgeblich die Grundlagen für die Datenerfassung und -verarbeitung mitgestaltet. Dies umfasste zunächst beratende Tätigkeiten bei der Definition der Anforderungen an die Patientenrekrutierung, um sicherzustellen, dass die gesammelten Daten klinisch relevant und repräsentativ für die angestrebten KI-Ziele sind.

Ein notwendiger Schritt für die erforderliche Datenqualität ist eine möglichst detallierte und fehlerfreie Annotation der Audiosignale. Zu diesem Zweck haben wir mehrere Annotations-Tools implementiert, erprobt und dem gesamten Projektteam vorgestellt. Nach dieser umfassenden Evaluierung fiel die Wahl auf *Labelstudio*, da es die Anforderungen des Projekts am Besten erfüllte. Das Programm bietet zentrale Funktionen zur

Qualitätssicherung und Nachvollziehbarkeit. Es protokolliert automatisch eine vollständige Versionshistorie aller Änderungen und ermöglicht das parallele Labeln derselben Audiodateien durch mehrere Nutzer. Diese Vorgehen des redundanten Labelings gewährleistet die Validität der Annotationen.

Darüber hinaus wurde eine individuelle Anpassung des Labelstudio-Interfaces an die vorgegebene Datensatzstruktur umgesetzt. Die annotierende Person kann nun nicht nur eine Gesamtdiagnose vergeben, sondern auch spezifische Zeitbereiche gesondert markieren. Diese Fähigkeit zur detaillierten Kennzeichnung spezifischer Geräusche im Signal ist essenziell, da sie in späteren Analysen die Grundlage für eine tiefere und genauere Auswertung der Daten schafft. Die angepasste Benutzeroberfläche wird in Abbildung 2.4 veranschaulicht.

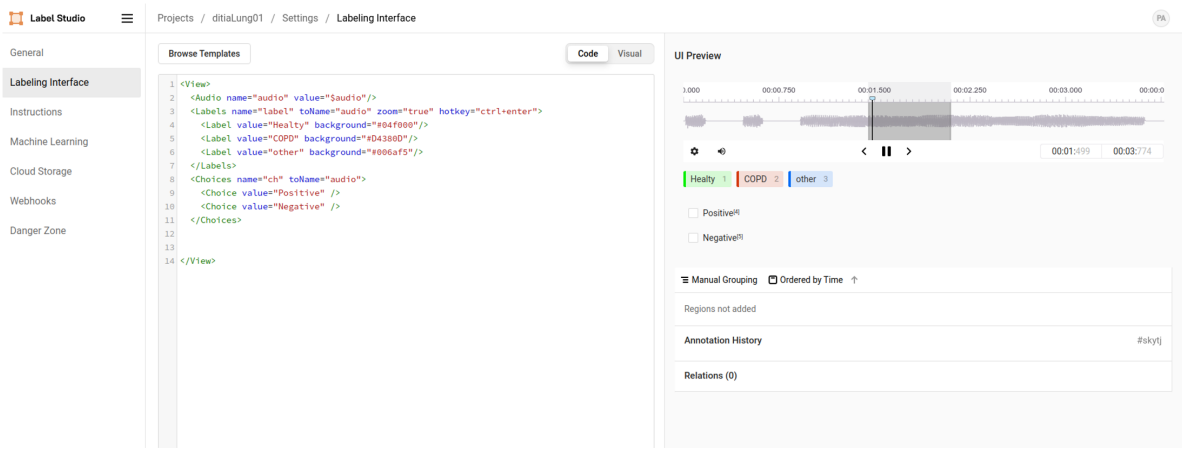


Abbildung 2.1: Interface von Labelstudio

2.1.2 Vorbereitende Datenanalyse (AP4)

Dieses Kapitel beschreibt zunächst den verwendeten Jordan-Datensatz und analysiert dessen spezifische Qualitätseigenschaften. Anschließend wird eine Methode zur erlernten Rauschreduzierung präsentiert, dessen Ziel es ist, die relevanten akustischen Merkmale im Signal gezielt von verfälschenden Störanteilen zu bereinigen.

Datensatz

Grundlage für die Entwicklung des digitalen Diagnosesystems sind die Daten aus der Patientenstudie. Dieser selbst erhobene Datensatz soll bestehende Einschränkungen öffentlicher Datensätze verbessern. Dazu haben wir im Projektteam die notwendigen Spezi-

fikationen erarbeitet, um eine ideale Datenbasis für unsere Machine-Learning-gestützten Verarbeitung zu schaffen. Dieser definierte Datensatz stellte einen deutlichen Mehrwert gegenüber bestehenden Datensätzen dar, da er einen signifikant tieferen Detailgrad und eine höhere Annotationstiefe vorsieht. Das technische Konzept umfasste zehn einheitliche Messkanäle zur Gewährleistung standardisierter Aufnahmen. Vor allem aber wurde eine umfassende Patienten-Auskultation mit über 100 Annotationen pro Patient spezifiziert, welche weit über reine Diagnosen hinausgeht. Dazu zählt die Erfassung spezifischer, klinisch hochrelevanter Merkmale wie Hustenmuster, Rauchersymptome und das Auftreten von Geräuscharten, die laut Fachärzten für eine Krankheit sprechen. Diese neuartigen Informationskategorien und der hohe Detailgrad der Annotationen sind in bestehenden Datensätzen nicht verfügbar.

Aufgrund von fortlaufenden Verzögerungen in der Erstellung des Datensatzes wurde die Entscheidung getroffen, die KI-basierte Diagnosesoftware auf einem bestehenden, öffentlichen Datensatz zu entwickeln und auszuwerten. Zusammen mit der MHH haben wir uns dabei auf den Jordan-Datensatz geeinigt. Im Rahmen einer Qualitätsuntersuchung des Jordan-Datensatzes haben wir jedoch einige Eigenschaften festgestellt, die die Aufgabe der digitalen Diagnose erschweren. Wie wichtigsten Erkenntnisse dazu sind im folgenden aufgelistet:

1. Mangelnde Annotationstiefe: Die Annotierungen des Jordan-Datensatzes sind sehr rudimentär. Abgesehen von den grundlegenden biometrischen Angaben wie Körpergröße und Gewicht, sind kaum Patienteninformationen verfügbar. Ursprünglich war eine deutlich höhere Datentiefe vorgesehen, inklusive Informationen zu Vorerkrankungen oder krankheitstypischer Geräuschmerkmale in der Audioaufnahme.

2. Eingeschränkte Diagnose: Die annotierte Diagnosevielfalt umfasst lediglich die Klassen: Gesund, Herzfehler, Asthma, COPD, Pneumonie und Lungen-Fibrose. Dies ist ein deutlich verminderter Detailgrad gegenüber der ursprünglichen Planung von bis zu 13 Lungen-spezifischen Diagnoseklassen.

3. Inhomogenität der Daten: Der Jordan-Datensatz nutzt nur ein Mikrofon pro Aufnahme (unimodal statt multimodal), welches zusätzlich mit unterschiedlichen Stethoskoptypen und aufgenommen wurde. Diese Aufnahmevariabilität geht mit teils inkonsistenten Lautstärken und Rauschanteilen einher, welches die Diagnose deutlich erschwert.

4. Unausbalancierter Datensatz: Der Datensatz weist ein unausgewogenes Klassenverhältnis auf. Die Klasse Asthma ist im Vergleich zu den anderen Pathologien (COPD, Pneumonie und Lungenfibrose) deutlich überrepräsentiert, siehe Tabelle 2.1

Eine weitere Herausforderung in Bezug auf den Jordan-Datensatz war die Erstellung eines Trainings- und Test-Split, welches eine Voraussetzung für datengetriebene KI-

Nr.	**Klasse**	**Train Dataset**	**Test Dataset**
1.	Gesund	64	41
2.	Asthma	59	37
3.	Herzinsuffizienz	33	21
4.	COPD	17	10
5.	Penumonie	10	5
6.	Lungenfibrose	8	4
	Summe	191	118

Tabelle 2.1: Anzahl der Aufnahmen pro Klasse bei einem genutzten Trainings/Test-split von 60/40.

Modelle ist. Eine vorgegebene Datenaufteilung ist im Jordan-Datensatz nicht vorhanden. Deshalb haben wir zusammen mit der MHH ein Programm zur automatisierten Erstellung eines Train/Test-Split entwickelt. Dabei ist methodisch sichergestellt, dass sowohl die Klassenbalance der Zielkrankheiten als auch wichtige Zusatzmerkmale wie Körpergröße und Gewicht gleichmäßig in Trainings- und Testdatensatz enthalten sind. Dies minimiert Verzerrungen und gewährleistet eine hohe Validität der späteren Modellaussagen. Die resultierende Aufteilung des Datensatzes ist in Tabelle 2.1 dargestellt. Für unsere Anwendung haben wir 40% der Daten für das Testset verwendet, um eine ausreichend große Anzahl an Aufnahmen im Testszenario zu gewährleisten. Perspektivisch kann dieses automatisierte Tool auch dazu genutzt werden den Datensatz aus der Patientenstudie nach gegebenen Fairnessbedingungen aufzuteilen, da sich die genutzten Entscheidungsmerkmale der beiden Datensätze überschneiden.

Denoising Autoencoder (DAE)

Um das verrauschte Eingangssignal aufzuarbeiten und nicht notwendige Signalanteile zu filtern, wurde ein Denoising Autoencoders (DAE) implementiert. Das Roh-Signal in der Auskultation stellt eine Mischung aus dem relevanten Lungensignal sowie Störkomponenten aus Herzsignal und Umgebungsrauschen. Die Aufgabe des DAE-Modells ist es, diese Störanteile herauszufiltern und so das reine Lungensignal zu rekonstruieren. Zu diesem Zweck wurde ein virtueller Datensatz erstellt, der künstlich durch die Überlagerung von 'reinen' Herz- und Lungengeräuschen mit Rauschsignalen generiert wird. Das DAE-Modell ist darauf trainiert worden, aus dem gemischten Signal, das 'reine' Lungensignal zurückzugewinnen.

Derzeit befindet sich der Denoising Autoencoder in der initialien Trainingsphase mit nur wenigen Datenbeispielen. Grund dafür ist, dass diese Methode ausgiebige, manuelle Anpassungen des virtuellen Datensatzes an die spezifischen Rauscheigenschaften des

Zieldatensatzes erfordert. Da der geplante Datensatz aus der Patientenstudie aktuell nicht verfügbar ist, konnte nur das DAE-Grundgerüst sowie die Logik zur automatisierten Generierung des virtuellen Datensatzes implementiert werden. Die abschließende Anwendung und das Training auf den finalen Projektdaten steht noch aus.

2.1.3 Datenverarbeitung und Maschinelles Lernen (AP5)

Dieses Kapitel konzentriert sich auf die wissenschaftlich-technische Realisierung der Diagnosesoftware. Im Zentrum steht die Entwicklung eines KI-gestützten Audioklassifikators, der Lungenkrankheiten diagnostiziert. Wir präsentieren die Vorgehensweise in drei systematischen Schritten: Zuerst wird die Datenaufbereitung als Fundament der Datenqualität erläutert, gefolgt von der Augmentierung zur Erweiterung der Datenbasis und schließlich die detaillierte Vorstellung des Digitalen Diagnosemodells.

2.1.4 Datenaufbereitung

Um eine hohe Modellperformance zu erreichen, schafft die Datenaufbereitung die methodische Basis, indem sie die Audiosignale für das maschinelle Lernen aufbereitet und die wichtigsten Merkmale für die Diagnose extrahiert.

Obwohl tracheale Atemgeräusche laut Koehler et al. [2] bis 2000 Hz wahrnehmbar sind, zeigten unsere empirischen Analysen, dass die niederfrequenten Signalanteile bis 500 Hz die höchste diagnostische Aussagekraft für die Klassifikation von Lungenkrankheiten besitzen. Aus diesem Grund wurde entschieden, die Audiosignale mittels einer Tiefpassfilterung auf 500 Hz zu begrenzen. Diese gezielte Reduktion des Frequenzspektrums ermöglicht es, eine höhere Auflösung im relevantesten Frequenzbereich zu erzielen und somit die Erkennung der Krankheitsklassen zu optimieren.

Im nächsten Schritt der Datenaufbereitung wird das bereinigte Audiosignal mittels eines Mel-Spectogramms in den Frequenzbereich transformiert. Das Spektrogramm visualisiert die Frequenzverteilung über die Zeit, wobei die Mel-Skala eine biologisch plausible Skalierung nutzt, die sich besser an der menschlichen Hörempfindung und somit an den relevanten Audiocharakteristika orientiert.

Augmentierung

Um die Generalisierungsfähigkeit des Diagnosemodells trotz begrenzter Datenbasis zu erhöhen, wurden umfangreiche Datenaugmentierungstechniken implementiert. Diese zielen darauf ab, realistische Datenvariationen zu generieren. Die angewandten Methoden umfassten das *Random Clipping*, bei dem zufällige zeitliche Ausschnitte zur lokalen

Merkmalserkennung genutzt werden, sowie *Speed Perturbation* und *Pitch Shift*, um natürliche Schwankungen in Geschwindigkeit und Tonhöhe zu simulieren und das Modell robuster gegenüber Frequenzverschiebungen zu machen. Zusätzlich wurde *Random Noise* verwendet, um die Robustheit gegenüber verrauschten realen Anwendungsszenarien zu verbessern.

Digitales Diagnosemodell

Die Kernkomponente zur diagnostischen Auswertung ist das Diagnosemodell. Das Modell nutzt das bereits definierte Mel-Spektrogramm als primäre Eingabe, um die Klassifikation des Signals in die entsprechenden Krankheitskategorien durchzuführen.

Zur Identifizierung der leistungsfähigsten Lösung wurden zwei unterschiedliche Arten von Modellen untersucht:

1. Die etablierte, CNN-basierte ResNet-Architektur [1], eignet sich besonders für die lokale Merkmalsextraktion in bildhaften Daten.
2. Der moderne, Transformer-basierte Ansatz Audio Spectrogram Transformer (AST), erfasse mithilfe des Attention-Mechanismus globale Abhängigkeiten über das gesamte Spektogramm [7].

Da der Transformer-basierte Ansatz AST einen deutlich höheren Speicherbedarf erfordert, musste die Batchsize von 128 auf 64 für diesese Modelle reduziert werden. Beide Modellarten wurden anschließend, mittels des Cross-Entropy-Losses [5] über 300 Epochen optimiert.

In Tabelle 2.2 sind die Ergebnisse der Modellevaluation aufgelistet. Dabei wurden für den CNN-Ansatz zwei verschiedene Modellgrößen, ResNet18 und ResNet50, sowie für den Transformer-basierten Ansatz die Modellgrößen Tiny und Small untersucht.

Model	**Accuracy**	**Specificity**	**F1-Score**	**Recall**
Resnet18	66.4	93.2	67.0	69.3
Resnet50	63.3	93.0	64.4	63.8
Ast-Tiny	50.3	87.9	44.9	33.0
Ast-Small	44.1	88.2	46.6	40.8

Tabelle 2.2: Qualitative Ergebnisse über verschiedene CNN und Transformer Modelle. Als Metrik ist die Accuracy, die Spezifität, F1-Score und der Recall aufgeführt

Um die Modelle objektiv zu bewerten, nutzten wir die folgenden Metriken: Die Accuracy (Genauigkeit) zeigt wie zuverlässig das Modell in allen Fällen richtig liegt, indem sie den

prozentualen Anteil aller korrekten Vorhersagen misst. Die Spezifität (Specificity) gibt an, wie gut das Modell gesunde Diagnosen korrekt erkennt. Im Gegensatz dazu misst der Recall, wie viele der tatsächlich vorhandenen Krankheiten das Modell erfolgreich identifiziert hat. Der F1-Score dient als ausgewogenes Verhältnis zwischen dem Recall und der Precision (Präzision), um die Qualität der Vorhersagen fair zu beurteilen.

Tabelle 2.2 zeigt, dass die CNN-basierten Architekturen im Vergleich zu den Transformer-Modellen eine höhere Klassifikationsgüte aufweisen. ResNet18 erreicht die höchste Performanz und erzielt die besten Werte mit einer Accuracy von 66,4% und einem F1-Score von 67,0%. Besonders hervorzuheben ist die hohe Spezifität von 93,2%, was impliziert, dass das Modell gesunde Diagnosen mit sehr hoher Zuverlässigkeit klassifiziert. Im Gegensatz dazu erreichen die Transformer-basierten AST-Modelle (Tiny und Small) eine signifikant geringere Klassifikationsgüte. Insbesondere die niedrigen Recall-Werte von lediglich 33,0% respektive 40,8%, die sich nicht im Verhältnis zur Accuracy verhält, indizieren einen starken Bias der Modelle zugunsten der überrepräsentierten Klassen. Dies deutet darauf hin, dass die evaluierten Transformermodelle weniger die gewünschten spezifischen akustischen Merkmale (Geräusche) zur Diagnose nutzen, sondern stattdessen primär lernen, die Klassenverteilung des Datensatzes zu widerspiegeln. Basierend auf den gemessenen Metriken wird ResNet18 als die effektivste Architektur für die vorliegende diagnostische Aufgabe ausgewählt.

Qualitätsanalyse des Jordan-Datensatzes

Die Entwicklung eines zuverlässigen Diagnosemodells hängt maßgeblich von der Qualität der Trainingsdaten ab. Aus diesem Grund ist eine detaillierte Analyse der datenspezifischen Eigenschaften des Jordan-Datensatzes notwendig. Diese Untersuchung dient dazu, die für die Klassifikation von Lungenkrankheiten relevanten akustischen Merkmale zu identifizieren und sie von jeglichen Störsignalen abzugrenzen.

Im Zuge dieser Analyse wurden drei Hauptkategorien von Fehlertypen identifiziert, welche die Modellperformanz negativ beeinflussen. Diese Mängel demonstrieren eine deutliche Abweichung vom angestrebten idealen, klaren Atemgeräusch (wie es exemplarisch in Teilbild a dargestellt ist). Die wichtigsten dieser Fehlertypen sind in Abbildung 2.2 visualisiert und werden im Folgenden beleuchtet:

1. Hohe Varianz der Lungenlautstärke (Teilbild b): Ein hauptsächliches Problem ist die inkonsistente Lautstärke der Atemgeräusche. Wie in Teilbild b demonstriert, sind viele Aufnahmen so leise, dass die respiratorische Aktivität akustisch kaum noch feststellbar ist und das Grundrauschen dominiert. Diese unkontrollierte Variabilität erschwert die Modellerstellung erheblich. Sie entsteht, da der Jordan-Datensatz keine standardisierten Vorgaben für die verwendeten Mikrofontypen und die Messpositionen festlegt. Es ist jedoch zu erwarten, dass dieser Fehlertyp

im Zieldatensatz dieses Projekts deutlich reduziert wird, da hier ein einheitlicher Mikrofontyp zum Einsatz kommt und die multimodalen Messpositionen vereinheitlicht werden.

2. Geräuschartefakte durch Hardware und Umgebung (Teilbild c): Unerwünschte Artefakte (z. B. durch Mikrofonreibung) und Umgebungsgeräusche sind typische Fehlerquellen im Jordan-Datensatz, die laut und deutlich im Signal präsent sind. Diese externen Geräusche reduzieren die Verwertbarkeit des Signals und die Zuverlässigkeit der Diagnose. Das für dieser Patientenstudie verwendete Stethoskop verfügt über eine destruktive Mikrofonarchitektur, in der Umgebungsgeräusche mittels eines zweiten Mikrofons mitgemessen werden um diese auf analoger Ebene zu entfernen. Daher ist zu erwarten, dass diese Geräuschartefakte nicht mehr im Zieldatensatz vorhanden sind.

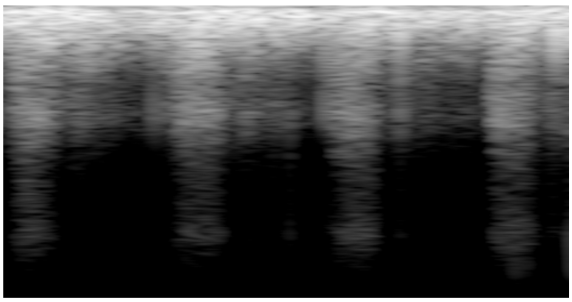
3. Dominanz kardialer Geräusche (Teilbild d): Ein weiteres Problem stellt die Überlagerung des Signals durch laute Herzgeräusche dar, die vor allem in den oberen Frequenzbereichen messbar sind. In den betroffenen Aufnahmen überwiegt das Herzsignal gegenüber dem eigentlichen Lungengeräusch. Dieses unerwünschte Herz-Artefakt soll jedoch im geplanten Zieldatensatz reduziert werden, indem zehn multimodale Mikrofone mit standardisierten Messpositionen verwendet werden. Konkret bedeutet dies: Messpositionen, die weiter vom Herzen entfernt liegen, weisen einen höheren Dämpfungsgrad für die Herzgeräusche auf, wodurch nach dem störende Artefakt effektiv gefiltert werden kann.

Zusammenfassend weist der Jordan-Datensatz eine Vielzahl an Fehlertypen auf, welche die angestrebte Klassifikationsgenauigkeit des Diagnosemodells signifikant beeinträchtigt. Im Rahmen der Qualitätskontrolle während der Partientenstudie sind diese Kriterien regelmäßig zu prüfen. Aufgrund der anhaltenden Verzögerungen bei der Erstellung des eigenen Datensatzes wird für den Machbarkeitsnachweis der digitalen Diagnosesoftware weiterhin der Jordan-Datensatz genutzt.

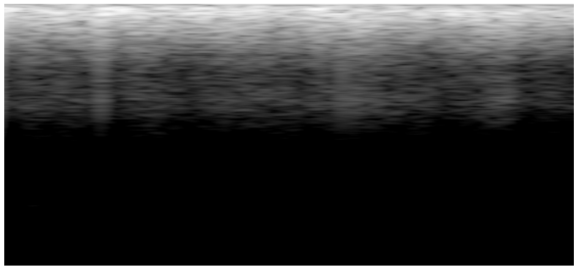
2.1.5 Erweiterte ML-Untersuchungen (AP9)

Ergänzend zu den reinen Klassifikationsergebnissen aus der Modellevaluation (Kapitel 2.1.4), welche die Leistungsfähigkeit des digitalen Diagnosemodells belegen, wurde das System um einen Post-Hoc-Ansatz aus dem Bereich der Interpretierbaren Künstlichen Intelligenz (Interpretable AI) erweitert.

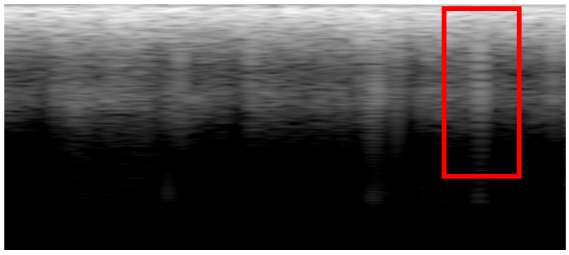
Zu diesem Zweck integrierten wir die Methode Grad-CAM [3] in das bereits bestehende Framework. Grad-CAM erzeugt eine visuelle Wahrscheinlichkeitsverteilung über die Bestandteile des Eingangssignals, die Aufschluss darüber gibt, welche spezifischen



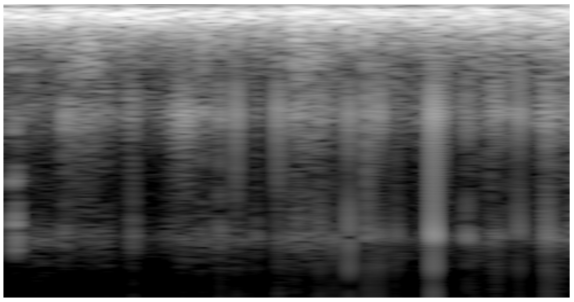
(a) Normales Atemgeräusch



(b) leises Atmen



(c) Mikrofonreibung



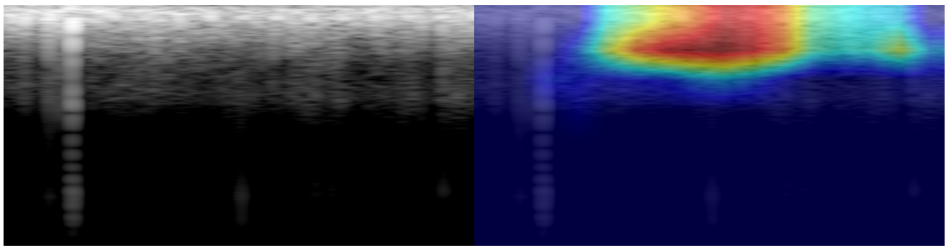
(d) laute Herzgeräusche

Abbildung 2.2: Visualisierung einiger ungewollter Signaleigenschaften im Jordan-Datensatz. Dazu gehören: zu leises Atmen (b), Mikrofonreibung (c) am Körper und zu lautes Atmen (d).

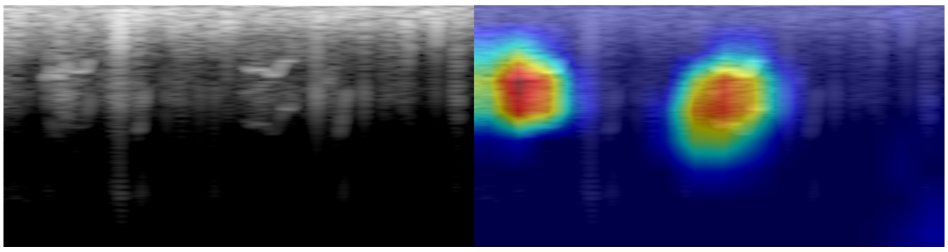
Frequenz- und Zeitabschnitte des Audiosignals die Aktivität einer bestimmten Krankheitsklasse am stärksten beeinflusst haben. Wird beispielsweise die Diagnose Asthma gestellt, kann mittels Grad-CAM präzise nachvollzogen werden, welche Signalabschnitte für die Klassenzugehörigkeit ausschlaggebend waren. Das übergreifende Ziel dieser Erweiterung ist es, das Vertrauen in das Digitale Diagnosemodell zu stärken und seine Entscheidungen für den Anwender transparent und nachvollziehbar zu machen.

Ein exemplarisches Ergebnis der Grad-CAM-Methode ist in Abbildung 2.3 visualisiert. Die Darstellung zeigt die primären Modellaktivierungen für eine gesunde Aufnahme (a) und eine Asthma-Diagnose (b). Jede Abbildung präsentiert auf der linken Seite das reine Mel-Spektrogramm und auf der rechten Seite die entsprechende Heatmap der neuronalen Aktivierungen. Die Achsen spiegeln dabei den Aufbau des Spektrogramms wider: Die x-Achse bildet den zeitlichen Verlauf ab, während die y-Achse die Frequenzen darstellt, wobei die niedrigen Frequenzbereiche oben liegen.

Die gesunde Aufnahme (Signal a) zeigt eine hohe Aktivierung primär im niederfrequenten Bereich. Zudem ist am Beginn dieser Aufnahme ein Mikrofonreibegeräusch zu erken-



(a) Gesund



(b) Asthma

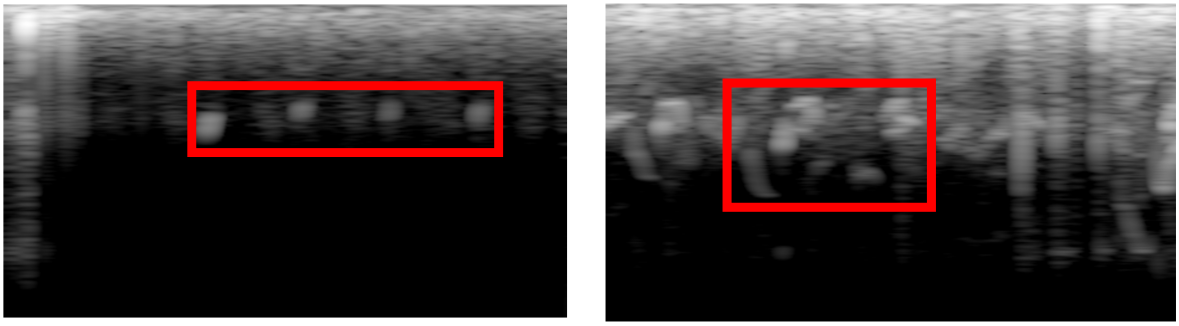
Abbildung 2.3: Visualisierung der Grad-CAM Ergebnisse zwischen Klasse Gesund (a) und Asthma (b).

nen (durch eine vertikal gepunktete Linie markiert). Das Modell hat dieses Störgeräusch als nicht relevant für die Klassifikation erkannt. Im Gegensatz dazu zeigt Signal b einen korrekt diagnostizierten Fall der Klasse Asthma. Hier hat das Modell für seine Entscheidung gezielt die Bereiche mit der markanten Wellenform als ausschlaggebend gewertet, um die Diagnose Asthma zu stellen.

Weitere Beispiele für Geräuschartefakte der Klasse Astma, die von Grad-CAM als relevant erkannt wurden, sind in Abbildung 2.4 dargestellt. Die jeweiligen, vom Modell als bewerteten Geräusche sind in den Darstellungen jeweils mit einer roten Box markiert.

2.1.6 Sonstige Arbeitspakete

Die Arbeitspakete AP 1 (Sensorik und Aufnahmesystem) und AP 2 (Forschungssoftware für Datenerfassung) wurden in enger Kooperation mit den übrigen Projektbeteiligten bearbeitet. Wir unterstützten das Team durch die Mitarbeit an den Geräteanforderungen und Spezifikation des Aufnahmesystems die notwendig sind um die Qualität des späteren Datensatzes sicher zu stellen. Der Austausch mit den Projektteams fand durch regelmäßige Abstimmungen statt, wobei ebenfalls beratende Tätigkeiten hinsichtlich der Softwareanforderungen geleistet wurden. Ein zentraler Inhalt dieser Zusammenarbeit war die Optimierung des Labeling-Prozesses: Um den manuellen Aufwand zu verringern, musste beispielsweise sichergestellt werden, dass die Labeling-Annotationen effizient in die



(a) Geräusch 1 **(b)** Geräusch 2

Abbildung 2.4: Visualisierung der Modellaktivierung für Asthma-Artefakte (Grad-CAM). Die Rote Box markiert den Bereich, den das Modell vorzugsweise als Feature für die klasse Asthma identifiziert

Datenbankstruktur aufgenommen werden konnten.

Innerhalb von AP 6 (Versuchsgeräte) wurde eine detaillierte Signalanalyse des Rohsignals durchgeführt, um die Grenze der relevanten Frequenzanteile zu bestimmen, bei der das rekonstruierte Signal keinen qualitativen Verlust an Geräuschen aufweist. Anhand von Beispielen aus dem Jordan-Datensatz konnte gezeigt werden, dass Frequenzanteile oberhalb von 1500 Hz vom menschlichen Gehör nicht mehr als Verschlechterung der Signalqualität wahrgenommen werden. Diese Erkenntnis ist relevant, da die empirischen Studien des digitalen Diagnosemodells (Kapitel 2.1.4) bereits belegten, dass die wichtigsten Klassifikationsmerkmale bis zu einer Frequenz von 500 Hz am besten funktionieren. Die Frequenzanteile zwischen 500 Hz und 1500 Hz sind zwar für den Menschen hörbar, stellen für die maschinelle Klassifikation jedoch keine klassen-diskriminierende Merkmale dar.

2.2 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Im Rahmen des Projekts DigitaLung ist uns gelungen, ein digitales Diagnosesystem für Lungenerkrangungen zu entwickeln. Durch die Nutzung des öffentlich verfügbaren Jordan-Datensatzes konnten wir demonstrieren, dass sich gängige Lungenpathologien wie Asthma, COPD (Chronisch-obstruktive Lungenerkrankung) und Pneumonie zuverlässig anhand von Audioaufnahmen der Lungengeräusche klassifizieren lassen.

Ein zentraler Bestandteil unserer Entwicklung war zudem die Einführung eines interpretierbaren KI-Modells. Dieses Modell geht über die reine Klassifikation hinaus, indem

es Geräuschbestandteile identifiziert, welche spezifisch auf die diagnostizierte Krankheit hindeuten. Dies liefert behandelnden Ärzten eine transparentere Entscheidungsbegründung und erhöht damit die Akzeptanz und Verlässlichkeit des digitalen Auskultationssystems in der klinischen Praxis.

Im Rahmen des Gesamtprojektes sollte ein Lungendatensatz (AP3) mithilfe des neu entwickelten digitalen Stethoskops erstellt werden. Ursprünglich war diese Projektphase bis PM 30 geplant, jedoch hat sich dieses Ziel im Laufe des Projekts fortlaufend verschoben. Demzufolge mussten einige darauf aufbauende Tätigkeiten wie beispielsweise die Datenvorverarbeitung, die Datenanalyse, oder das geplante Training der Diagnosesoftware mehrfach verschoben werden.

Trotz dieser Verzögerungen erbrachte die Leibniz Universität Hannover (LUH) notwendige und angemessene Beiträge zur Sicherstellung der Datenqualität und der algorithmischen Entwicklung. Dazu gehörte die beratende Funktion bei der Datensatzerstellung und die Integration eines geeigneten Labeling-Tools, um die spätere maschinelle Verwertbarkeit der Daten sicherzustellen. Parallel dazu wurde ein KI-Diagnosesoftwaremodul entwickelt und auf dem Jordan-Datensatz angewendet. Diese Entscheidung erlaubte es uns, auch ohne den eigentlichen Zieldatensatz das entwickelte Modell zu evaluieren. Es ist festzuhalten, dass der genutzte Jordan-Datensatz die Evaluierung aufgrund mehrerer Limitationen erschwerte. Es fehlte an ausreichender Datenmenge und einer standardisierten Aufnahmequalität, da unterschiedliche Geräte und Messpositionen verwendet wurden. Ungeachtet dieser Schwierigkeiten gelang es uns, die Robustheit der genutzten Algorithmen zu demonstrieren und ein funktionierendes digitales Diagnosesystem zu entwickeln.

2.3 Voraussichtlicher Nutzen

Die Ergebnisse führen zum erfolgreichen Nachweis der Funktionsfähigkeit eines innovativen Diagnosesystems, das einen substanziellen Nutzen für die Erkennung von Lungenkrankheiten bietet. Das Kernergebnis ist die Entwicklung eines digitalen, KI-basierten Diagnosesystems, welches einen maßgeblichen Fortschritt in der Differentialdiagnose von Atemwegserkrankungen darstellt. Dieses System ist in der Lage, aus akustischen Lungenaufnahmen zuverlässig spezifische Krankheiten wie Asthma, COPD oder Lungenpneumonie zu erkennen. Eine wesentliche Erweiterung des Modells liefert interpretierbare Erklärungen zur Entscheidungsfindung des Modells, die für die Diagnose ausschlaggebend sind. Dieser Post-Hoc-Ansatz erhöht die Transparenz der Systementscheidungen, was für die Akzeptanz und den Einsatz in der medizinischen Praxis von großer Bedeutung ist. Durch diese Transparenz können Ärztinnen und Ärzte fundierter in ihrer Lungendiagnose unterstützt werden, indem sie nicht nur ein Klassifikationsergebniss,

sondern auch die zugrunde liegenden Merkmale erhalten. Die Entwicklung dieses intelligenten Diagnosesystems trägt somit direkt zur Verbesserung der Gesundheitsversorgung bei und demonstrieren das Potenzial von KI-basierten Lösungen im Medizinbereich.

2.4 Erfolgte Veröffentlichungen

Für die Veröffentlichung aussagekräftiger und valider Erkenntnisse fehlen uns die notwendigen Daten der Studie zur Lungenauskultation. Eine Konferenzpräsentation war unter diesen Umständen bisher nicht möglich.

Literaturverzeichnis

[1] Kaiming He et. al. Deep residual learning for image recognition, 2015.

[2] Koehler U. et. al. Atemgeräusche und atem-nebengeräusche, 2016.

[3] Ramprasaath R. Selvaraju et. al. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2019.

[4] Sovijärvi et. al. Computerized respiratory sound analysis (corsa): Recommended standards for terms and techniques, 2000.

[5] Zhilu Zhang et. al. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.

[6] P Forgacs. The functional basis of pulmonary sounds, 1978.

[7] James Glass Yuan Gong, Yu-An Chung. Ast: Audio spectrogram transformer, 2021.